DOCUMENT RESUME

ED 193 631 CS 205 606

AUTHOR Humes, Ann
TITLE A Method for Evaluating Writing Samples. Technical Note.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
REPORT NO SWRL-TN-2-80/02
PUB DATE Apr 80
NOTE 32p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Evaluation Criteria; Evaluation Methods; Measurement Techniques; *Scoring; *Writing (Composition); *Writing Skills
IDENTIFIERS Holistic Approach; *Writing Evaluation

ABSTRACT

The discrepancy in the application of terminology in labeling systems for scoring writing samples is discussed in this paper and current scoring systems that exclude frequency counts of writing features are described. Two-thirds of the paper discusses and gives examples of general impression, marking, analytic scale scoring, and primary trait scoring. The last third of the paper compares and contrasts a new model for scoring developed by the Southwest Regional Laboratory (SWRL) with the three widely used systems and lists the following benefits that result from scoring writing samples with the SWRL system: (1) more efficient method of eliciting and evaluating writing samples administered to determine students' writing competency; and (2) greater ease in using diagnostic information to identify instructional inadequacies and assure student learning. The results of a pilot study using the SWRL model are referred to in support of the appropriateness of the design and the range of scoring. (AEA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193631

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SOUTHWEST REGIONAL LABORATORY TECHNICAL NOTE

DATE: April 15, 1980

NO: TN 2-80/02

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SWRL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A METHOD FOR EVALUATING WRITING SAMPLES

Ann Humes

ABSTRACT

The discrepancy in labeling systems for scoring writing samples is discussed and current scoring systems are described. Then a new model for scoring is presented by comparing and contrasting it with widely used systems.

CS205606

This document is intended for internal staff distribution and use. Permission to reprint or quote from this working document, wholly or in part, should be obtained from SWRL, 4665 Lampson Avenue, Los Alamitos, California 90720

A METHOD FOR EVALUATING WRITING SAMPLES

Ann Humes

The scoring of writing samples is currently an important topic in the writing-research community. This topic was investigated in order to develop a format for scoring the writing samples included in competency tests. The current literature was reviewed for recommendations, especially for those pertaining to quick, reliable, and simple scoring. Then an appropriate and efficient model was developed for eliciting and scoring writing samples. This paper discusses, as background, the discrepancy in application of terminology used to reference scoring types in the literature, describes current scoring systems, and presents the SWRL model by comparing and contrasting it with the other systems.

BACKGROUND

Authorities generally agree that any assessment instrument for writing should include one or more writing samples (e.g., Brown, 1979). However, they disagree over which procedures are appropriate for scoring writing samples. One reason for the discord is that the exchange of ideas and information is hampered by variation in the terminology that labels scoring types. According to Cooper (1977), "Some uncertainty about terminology seems quite natural in an area in which the early research and field work has been so uncoordinated and in which we have not reached wide agreement on standard procedures" (p. 4).

In order to evaluate the arguments presented for and against various scoring types, the reviewer of literature on writing-sample scoring must

first recognize which scoring system is referenced when the literature refers to scoring procedures such as "holistic" and "analytic"--terms that are applied inconsistently. For example, Smith (1979) classifies scoring procedures as either "analytic" or "holistic" (p. 2). The Los Angeles County Division of Curriculum and Instructional Services (A Common Ground, 1978) also suggests that there are "two kinds of scoring: holistic or analytic" (p. 9). Annich (1979) ascribes these labels to this scoring dichotomy, explaining that "In analytic scoring, various components of writing are examined individually. . . . In holistic scoring, papers are read for a total impression or effect" (p. 11).

Conversely, Lloyd-Jones (1977) classifies tests as either "atomistic" or "holistic," saying that "Atomistic tests rely on the assessment of particular features associated with skill in discoursing, whereas holistic tests consider samples of discourse only as whole entities" (p. 33). Cooper (1977) defines holistic scoring as "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing" (p. 4). He adds that "Some holistic procedures may specify a number of particular features and even require that each feature be scored separately, but the reader is never required to stop and count or tally incidents of the feature" (p. 4). In consonance with his definition of "holistic," Cooper categorizes both analytic scoring and general impression scoring as "holistic;" however, general impression scoring is the system many people (including Smith and Annich, noted above) are referencing when they contrast holistic and analytic scoring.

Understanding the specific scoring systems requires going beyond the labels to a description of the systems themselves. However, only the "non-count systems" are described in the following discussion because frequency counts of features should not be included in systems designed for quick and simple scoring. Furthermore, frequency counts are more appropriate for measuring skills rather than content (e.g., Friedman and Fowler, 1979; Winters, 1979).

SCORING SYSTEMS

Cooper (1977) has the most comprehensive description of scoring systems. He delineates seven types of non-count scoring: essay scale, dichotomous scale, feature analysis, center-of-gravity response, general impression marking, analytic scale, and Primary Trait Scoring. The last three are the systems most frequently used, so they are subsequently discussed in detail. For the following brief descriptions of the remaining four types of non-count scoring, Cooper (1977) is referenced.

Using the essay scale entails ranking a piece of discourse by comparing it with model essays that exemplify each rating across the span of the scale. Thus essay scales set absolute standards which students' writing samples must meet. Essay scales were widely used during the 1920's and 1930's, but they are rarely used today because, according to Cooper, other scales are more efficient and reliable.

The dichotomous scale consists of a series of statements to be answered "yes" or "no." Cooper states that no reliability data are available on use of this scale with individuals because it was originally developed for comparing a group's pre- and post-test performance. He doubts that using dichotomous scales for individuals will yield reliable scores.

Feature analysis focuses in great depth on one particular aspect of a piece of discourse. Cooper's one example of this kind of evaluation is an instrument devised by Stahl (1974). This instrument measures structural characteristics of composition in a discourse describing a pictured house. The stimulus allows for different integrations of the details of the structure of that house. Each criterion included in the scale is scored in terms of most to least sophisticated type of structure. This scoring procedure is complex, potentially confusing, and limited in scope.

The center-of-gravity response, as described by Cooper, has feedback rather than scoring as its purpose. The evaluator reads the paper and summarizes his/her reaction to the discourse in four different ways (e.g., the evaluator quickly summarizes the main points; the evaluator summarizes the paper in a single sentence).

As noted above, the scoring procedures frequently discussed and most widely used are general impression (hereafter referred to as "holistic" in accordance with this usage in much of the literature, including literature referenced below), analytic, and Primary Trait Scoring.

Holistic Scoring

Holistic scoring involves a comparative evaluation of essays, as is performed by the Educational Testing Service (1976), the developers of the system. Readers are trained to read and rate papers in relation to other papers in a set. The papers are frequently scored on a scale ranging from one to eight (Mullis, 1975). Although a range of four may be used (ETS, 1976), some authorities believe that a range of eight or nine is optimal for making discriminations in the overall quality of writing

when comparison is the basis for the scores (A Common Ground, 1978). For an example of a holistic scoring scale, see Figure 1.

Despite the training period ETS requires for holistic scorers, a few authorities laud the procedure as so simple that evaluators can merely sort a stack of essays into eight piles (Hogan and Mishler, 1979). Odell and Cooper (1978) hypothesize that the scoring ease is a significant reason for the method's wide acceptance.

9-8	Papers that are clearly excellent. The top score, a score of 9, is reserved for that paper clearly above an 8. Both 9's and 8's are papers which develop the topic with excellent organization, content, and insight. They display facile use of language and mastery of mechanics.
7	A thinner version of the excellent paper, still impressive, cogent, convincing, but less well handled in terms of organization, insight, or language.
6-5	An above-average paper, but it may be deficient in one of the essentials mentioned above. It may be less mature in thought or less well handled in terms of organization, syntax, or mechanics. The 5 paper is a thinner version of the 6. The papers below 5 are lower half papers.
4-3	A lower half paper, which maintains the general idea of the writing assignment, shows some sense of organization, but is weak in content, thought, language facility, mechanics. It may distort the topic or fail to deal adequately with one important aspect of the topic. The 3 paper is a thinner version of the 4.
2	A paper that makes an attempt to deal with the topic but demonstrates serious weaknesses in content and coherence and/or in syntax and mechanics. It is unacceptable for most standards.
1	An on-topic response that has almost no redeeming quality. It may be very brief or very long, but will be scarcely coherent and probably full of mechanical errors as well.
0	Given to a blank paper or a non-topic response.

Figure 1. Example of holistic scoring scale.*

*A Common Ground, 1978, p. 11.

However, holistic scoring is frequently criticized because the only diagnostic information it offers is about the comparative quality of a paper--it does not indicate why a paper is a "3," or a "5," or a "7." A study reported by the California State Department of Education (An Assessment of the Writing Performance, 1977) also criticized the system for this inadequacy: The rating scale of 1-9 "provided no information about specific strengths and weaknesses" (p. 5). Los Angeles County (A Common Ground, 1978) and the National Assessment of Educational Progress (NAEP) have also decried this aspect of the scale. Although NAEP used holistic scoring with a range of 0-8 in their 1969-70 assessment (1970), the project group subsequently disavowed this system because test results could not be interpreted satisfactorily (Mullis, 1975).

Analytic Scales

Analytic scoring is another frequently used technique. The definitive analytic scale was devised by Diederich (1966 and 1974). This scale consists of "General Merit" categories (the content characteristics) and "Mechanics" (form skills such as punctuation, capitalization, and grammar).

An analytic scale consists of a list of the prominent features of a particular writing type (e.g., expository discourse, narrative discourse). Thus it is designed to accommodate all tasks for that discourse type, regardless of differences in the stimuli eliciting the writing samples. Although five numerical ratings can be checked, only three ratings are described for each feature of the scale--low, middle, and high. When

an essay is scored, the ratings on various features are tallied for subtotals of content and form scores and for a total writing score. See Figure 2 for Diederich's scale for scoring expository-writing tasks and Figure 3 for an example of the scoring guidelines that accompany the scale. See also Figures 4 and 5. Figure 4 exemplifies how the Diederich scale changes according to discourse type, here a personal narrative; it also displays a modification of the form-skills (mechanics) portion of the scale. Figure 5 is an example of the scoring guidelines for using the scale in Figure 4.

Many authorities advocate the analytic scale, asserting, for example, that it facilitates quick and efficient scoring (A Common Ground, 1978). Additionally, an analytic scale, according to Cooper (1977), "will guide and focus raters' scoring, insuring enough agreement to permit a reliable score to come from summed multiple ratings" (p. 16). He adds that this score "can be used for any purpose to which scores are ordinarily put in education: prediction, placement, exemption, growth measurement, program evaluation, or experimental or correlational research" (p. 16).

Research studies also support analytic scales. Smith (1979) compared holistic and analytic scales and concluded that analytic scales have greater potential for diagnosing students' writing strengths and deficits. She also asserts that "Analytic rating procedures aptly meet the requirements of competency based assessment programs which call for information on specific skill strengths and deficits" (p. 3). Pitts (1979) developed and experimentally employed a four-point analytic scale corresponding to the features of a combination of writing types (i.e., narrative/descriptive).

Topic _____ Reader _____ Paper _____

	Low		Middle		High	
Ideas	2	4	6	8	10	
Organization	2	4	6	8	10	
Wording	1	2	3	4	5	
Flavor	1	2	3	4	5	_____*
Usage	1	2	3	4	5	
Punctuation	1	2	3	4	5	
Spelling	1	2	3	4	5	
Handwriting	1	2	3	4	5	_____*
					Sum	_____

Figure 2. Diederich Scale**

General Merit

2. Organization

High. The paper starts at a good point, has a sense of movement, gets somewhere, and then stops. The paper has an underlying plan that the reader can follow; he is never in doubt as to where he is or where he is going. Sometimes there is a little twist near the end that makes the paper come out in a way that the reader does not expect, but it seems quite logical. Main points are treated at greatest length or with greatest emphasis, others in proportion to their importance.

Middle. The organization of this paper is standard and conventional. There is usually a one-paragraph introduction, three main points each treated in one paragraph, and a conclusion that often seems tacked on or forced. Some trivial points are treated in greater detail than important points, and there is usually some dead wood that might better be cut out.

Low. This paper starts anywhere and never gets anywhere. The main points are not clearly separated from one another, and they come in a random order--as though the student had not given any thought to what he intended to say before he started to write. The paper seems to start in one direction, then another, then another, until the reader is lost.

Figure 3. Scoring Guidelines for Organization, the second feature on the Diederich scale for expository writing (cf. Figure 2).***

*The blanks are for recording subtotal scores for both content and form.

**Diederich, 1974, p. 54.

***Diederich, 1974, p. 56.

Analytic Scale

Reader _____ Paper _____

	Low		Middle		High
I. General Qualities:					
A. Author's Role	2	4	6	8	10
B. Style or Voice	2	4	6	8	10
C. Central Figure	2	4	6	8	10
D. Background	2	4	6	8	10
E. Sequence	2	4	6	8	10
F. Theme	2	4	6	8	10
II. Diction, Syntax, and Mechanics:					____
A. Wording	1	2	3	4	5
B. Syntax	1	2	3	4	5
C. Usage	1	2	3	4	5
D. Punctuation	1	2	3	4	5
E. Spelling	1	2	3	4	5
					____
			Total		____

Figure 4. Analytic Scale for a Personal Narrative.*

D. Background

The setting of the action is detailed so that it seems to give the events a "real" place in which to happen.

High The action occurs in a well-detailed place that you can almost see.

Middle Sometimes the setting seems vivid and real; but sometimes the action is just happening, and you are not really aware of what the setting is.

Low The action occurs without any detailed setting. You see the action, but you cannot see it in a certain place.

Figure 5. Scoring Guidelines for Feature I, D, Background.**

*Cooper, 1977, p. 24.

**Cooper, 1977, p. 22.

She concluded that such scales are valid measures of writing performance and that readers can make highly reliable ratings after a minimal amount of training. Winters (1979) compared the use of holistic scoring, analytic scoring, and frequency counts to score students' essays. She concluded that the Diederich expository scale "appears to be the best system for distinguishing between high and lower performance writers . . ." (p. 20).

Analytic scales have critics as well as proponents. Cooper (with Odell, 1978), criticizes two underlying assumptions of the scale: (1) Criteria for rating good writing can be derived from sophisticated readers' perceptions of discourse types rather than from the specific writing task itself, and (2) the same criteria for judging a task elicited by one stimulus for a specific type of writing (e.g., expository) can be used to judge all tasks eliciting that writing type, even though the tasks are evoked by different stimuli.

Primary Trait Scoring

Cooper now advocates Primary Trait Scoring. Even while he was still a proponent of the Diederich scale, he admired Primary Trait Scoring as "the most sophisticated of the holistic [i.e., non-count] evaluation schemes . . . and potentially the most useful" (1377, p. 11).

Primary Trait Scoring (PTS) was devised by the National Assessment of Educational Progress (NAEP). NAEP is a project currently funded by the National Institute for Education, and its primary purpose is "to gather information about the degree to which educational goals in important learning areas such as writing are being met nationally, and to make this information available to the public . . . so that problem areas can be specifically identified, priorities established and progress

over time determined" (Klaus, Lloyd-Jones, Brown, Littlefair, Mullis, Miller, and Verity, 1979, p. 5).

The primary trait of a writing task is determined by the purpose and audience for a specific piece of discourse and is rated on how well that "trait"* is manifested. Each stimulus for a writing task is analyzed to identify the characteristics that effectuate that trait in the corresponding response--a writing sample; the characteristics (features) are unique to that stimulus, not applicable to other stimuli for the writing type. These characteristics become the scoring criteria for the writing sample. Thus the scoring criteria for a writing task with an explanatory purpose will include some type of orderly presentation of specific information because "organization is usually the basis of the primary trait for a writing task requiring explanation" (Mullis, 1975, p. 11).

Although the analytic scale for expository (explanatory) writing also includes "Organization," as well as "Ideas" (factual content) as features to be scored, its accompanying scoring guidelines do not describe the specific organization or the type of factual content pertinent to the exact writing task. The Primary Trait Scoring system does. For example (Mullis, 1975), the primary trait of a discourse that presents a set of directions (an expository task) would include as "subtraits" a sequential presentation--organization--of specific instructions--factual content. "Successful papers will have that trait; unsuccessful papers will not, regardless of how clever or well written they may be in other respects" (Mullis, 1975, p. 5).

---

*The use of the term "trait" (rather than "traits") is somewhat confusing because the singular form parallels the use of the term "feature" to describe each criterion on an analytic scale; however, "trait" as used by PTS refers to something like a "sum" of the important features, or "subtraits."

Thus the important distinction of PTS is that specific writing tasks are evaluated--papers are not evaluated in relation to other papers in the set, as in holistic scoring, nor solely in relation to the general writing type, as in the Diederich analytic model. Furthermore, a paper can be excellent in terms of language, structure, and syntax, yet be rated as a failure if it does not exhibit the primary trait.

The stimulus for a writing task evaluated by PTS must be highly structured to facilitate scoring, since a "multi-focused stimulus often evokes responses too dissimilar to be evaluated using common criteria" (Mullis, 1975, p.8); the "more structured the task, the less difficult the scoring, since the essays . . . will be more uniform in focus" (p. 9). The corresponding scoring guide that is written for each stimulus describes the exercise in terms of specific characteristics displayed by a successful writing sample (Klaus et al., p. 15). From this scoring, examiners can develop a profile of students' writing abilities for the large elements of communication such as the ability to arrange information coherently (Klaus et al., 1979). For examples of PTS writing sample stimuli, see Figure 6 and 8. For the corresponding guidelines for scoring the primary trait, see Figures 7 and 9. The guidelines in Figure 7 represent the early PTS approach; Figure 9 displays a more current model.

NAEP (Mullis, 1975) suggests that readers for Primary Trait Scoring should have previous English or classroom experience and should go through a two-stage training period. The first stage consists of an orientation to the procedure. In the second stage, readers practice evaluating

papers that have previously been scored. Then the prospective readers and the reader trainer discuss the scores until everyone concurs. NAEP additionally suggests that an apprentice scorer should keep rating practice papers until he/she can score consistently.

Because PTS evaluation focuses on the primary trait of the writing task, the scoring is "independent of attention to mechanical and grammatical features" (Klaus et al., 1979, p. 23). Thus when readers evaluate the primary trait, "Handwriting does not matter . . . Mechanics do not matter" (Klaus et al., 1979, p. 29). NAEP does score some samples for grammar and mechanics in order to compare these scores with performance scores on the primary trait. NAEP believes that while descriptive scoring of grammar and mechanics can provide some information about essays, "tallies of error rates do not go beyond describing the level of technical competence exhibited in the writing" (Klaus et al., 1979, p. 4). When mechanics are concurrently scored, they are scored as the secondary or tertiary trait.

Definitive studies have not compared PTS with other scoring systems (Odell and Cooper, 1978). However, NAEP has reported on the field testing of its assessment, and PTS is being used by some school districts for their writing assessments. The Hawaii Writing Assessment (Kadomoto, 1979), for example, uses PTS. This test accommodates writing aspects other than the primary trait by evaluating syntax as the secondary trait, and punctuation, capitalization, and grammar as the tertiary trait. This design, therefore, provides for three categories of diagnostic information: composing (from scores of primary trait evaluation), syntax (from secondary trait evaluation), and technical skills (from tertiary evaluation of punctuation, capitalization, and

Look carefully at the picture.

These kids are having fun jumping on the overturned boat. Imagine you are one of the children in the picture. Or if you wish, imagine that you are someone standing nearby watching the children. Tell what is going on as he or she would tell it. Write as if you were telling this to a good friend, in a way that expresses strong feelings. Help your friend FEEL the experience too. Space is provided on the next three pages.

Figure 6. PTS Writing Sample Stimulus*

*Mullis, 1975, p. 15.

ENTRY INTO WORLD OF PICTURE

0 = NO RESPONSE. NO FURTHER SCORING.

1 = NO ENTRY INTO THE IMAGINARY WORLD OF THE PICTURE.
Respondents write about children, the boat, or about anything else to do with the picture. However there is only a single statement or the information is too disjointed to make a point. Random details, bits of information, or lists of observations that do not create a situation. Also include pure description, papers that only report what's in the photograph or picture.

2 = ENTRY INTO THE IMAGINARY WORLD OF THE PICTURE.
Respondents have accepted the world of the picture. However, the control and consistency necessary to create a structured presentation are lacking. Often there is no structure to the world of the picture. This is evidenced by few internal transitions and details that don't harmonize. The ideas may be related, but don't make a whole. Lack of consistency and transitions result in little logical progression of ideas. On the other hand, these papers may have structure, but the narrative is not worked out. There is little imagination on the part of the writer to create the story. An attitude may be stated, but it's not illustrated. You do not "feel" a mood has been created. Other papers may have some structure and the outline of story or the hint of a mood, yet neither are developed. Generally "2" papers are either undeveloped or developed in a helter-skelter or confusing manner.

3 = GOOD ENTRY INTO THE IMAGINARY WORLD OF THE PICTURE.
Papers are generally competent. Respondents evidence control and consistency to create a structured presentation. Often there are strong topic sentences and good transitions. However, the good quality of the papers is marred by development that is skeletal or somewhat uneven. Narratives display imagination, but often are left with gaps or other uneveness [sic]. Details may be inadequate, excessive, or unclear. Attitudes may be stated and somewhat supported, but [sic] necessarily presented to "help your friend feel."
A definite mood is not created. The underdeveloped paper has a definite beginning and end yet there is not enough material to fill out the structure or it is contrived. The unevenly developed papers either have excessive details or the structure is oddly filled out.

4 = EMOTIVE AND CONSISTENT ENTRY INTO THE IMAGINARY WORLD OF THE PICTURE.
These papers are structurally whole. Loose ends have been tied up or cut off (although a strong paper without closure can be rated in this category). Papers are consistent. Narratives are well and evenly developed or attitudes are expressed so a definite mood is created. You do "feel" the experience. The structure is unified and supported by imaginative and evocative details.*

Figure 7. Scoring Guidelines for Primary Trait: Entry Into World of Picture.**

*Two additional categories provide for designating unscorable papers (as unreadable or as a sample that is not responding to the writing-sample stimulus) and another additional category is added to allow a scorer to say "don't know."

**Mullis, 1975, p. 18.

Sometimes people write just for the fun of it. This is a chance for you to have some fun writing.

Pretend that you are a pair of tennis shoes. You've done all kinds of things with your owner in all kinds of weather. Now you are being picked up again by your owner. Tell what you, as the tennis shoes, think about what's going to happen to you. Tell how you feel about your owner. Space is provided below and on the next two pages.

Figure 8. PTS Writing Sample Stimulus*

Primary Trait: Expression of feeling through elaboration of a role

1 = No established role. Respondent writes about tennis shoes or role is never related to any attributes or activities of tennis shoes (I went to the park). No indication of having an owner.

2 = No or little elaboration of role. Establish explicitly or implicitly the role of tennis shoes, but no distinctive character or personality emerges. Role may lapse, but should be evidence of a present or implied owner. Some types of "2" papers.

(1) only describe tennis shoes, ex. "I am a blue and yellow racing sneaker."
(2) only report experiences shared with owner, relationship with owner at surface level, ex. "We went to the park. Then we went to play tennis, then my owner took me home."
(3) only feelings, does not relate feelings to any experiences.
(4) inconsistent or contradictory role.

3 = Good elaboration of role. Expresses attitudes and feelings; develops shared experiences with owner or develops particular character and personality. May include reporting sections or some details irrelevant to personality or relationship with owner.

4 = Inventive and consistent elaboration of role. Use of vivid and creative details to sustain the role. Does not lapse out of role and includes few if any irrelevancies.

Other = Illegible, illiterate, misunderstands the task, writes on a totally different subject, I don't know, etc.

Figure 9. Scoring Guidelines for Primary Trait: Expression of Feeling through Elaboration of Role.**

*Klaus et al., 1979, p. 14.

**Klaus et al., 1979, p. 25.

grammar). Since scoring guidelines for secondary and tertiary traits are different from guidelines for the primary trait, they are exemplified in Figure 10.

PTS is also used for the Louisiana Writing Production test (Scherich, Liu, Kar, and Kettel, 1979), which has modified the primary trait concept by adding "syntactic acceptability" (p. 3). However, this characteristic should not be included in the scoring for a primary trait according to the PTS description of Klaus et al. (1979). The Louisiana Writing Production scores mechanical aspects of writing as the secondary trait. Readers are given "a considerable amount of training" (p. 3) and then are checked periodically during independent scoring to ensure consistency in ratings. Preliminary analysis indicates that consistency is high, with readers of eighth grade samples, for example, agreeing approximately 68% of the time.

Although the PTS system seems superior to many other non-count evaluation methods, its use for competency assessment can be criticized for several reasons. PTS does not evaluate grammar and mechanics except optionally as a secondary or tertiary trait; however, these features are commonly of interest to schools. Also questionable for competency testing is the PTS policy of assigning failure ratings to papers that are well written technically but do not manifest the primary trait as described in the scoring guidelines. Lloyd-Jones (1977) cites a potential problem when scoring must be done by classroom teachers rather than professional readers in his assertion that "In terms of convenience for the tester, Primary Trait Scoring is more difficult than other methods" (p. 40). Finally, PTS provides less specific diagnostic information than either analytic scales or the SWRL model discussed on the following page.

The secondary trait for this exercise is proper use of syntax. The responses will be scored according to the following levels of quality:

1. Secondary trait absent--Shows little or no sense of sentence structure.

2. Secondary trait present--Many sentences have structural weaknesses and grammatical errors.

3. Secondary trait adequate--Sentence structure generally correct. Some awkwardness, but does not detract from overall effectiveness.

4. Secondary trait excellent--Words and phrases combined effectively into permitted and appropriate language.

The tertiary trait for this exercise is proper use of punctuation, capitalization, and other conventions. The responses will be scored according to these four levels of quality:

1. Tertiary trait absent--So many errors that paper is difficult to read.

2. Tertiary trait present--Several errors that detract from overall effectiveness of writing.

3. Tertiary trait adequate--A few errors, but they do not detract from overall effectiveness.

4. Tertiary trait excellent--No errors.

Figure 10. Secondary and Tertiary Trait Guidelines.*

*Kadomoti, 1979, p. 2.

THE SWRL MODEL

The discussion above suggests that the scoring system used to evaluate writing samples for competency tests should (1) be easy to score, thus requiring little or no training of evaluators, (2) be scored for the specific content features of the writing task rather than for the general writing type elicited, and (3) provide diagnostic information. Additionally, form skills should be assessed because they are important to writing competency (Rentel, King, and Pappas, 1979); furthermore, performance on form skills may influence evaluators' scoring of content (either negatively or positively) if the scoring system does not provide for their separate evaluation.

The SWRL scoring system addresses these considerations. It combines the simplicity of the analytic scales, thus being easy to score, with the specificity of Primary Trait Scoring. Each highly structured stimulus also facilitates easy scoring because it is always accompanied by scoring criteria that pertain specifically to that stimulus. For an example of a writing sample stimulus see Figure 11. For the corresponding scoring key, see Figure 12.

As illustrated by the scoring key in Figure 12, the skills array of the scoring matrix consists of the features of good writing that the sample exhibits when a student employs the content and form skills appropriate for both the specific writing task and the grade level of assessment. As previously noted (Mullis, 1975), an explanatory writing sample that gives directions should sequentially present specific instructions. Also, this information should be conveyed in precise language. The

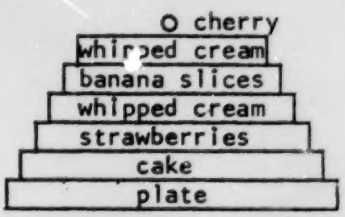
Directions: Read the contents and study the layers of the Fruit and Whipped Cream Treat.

Fruit and Whipped Cream Treat

Contents

1 yellow cake
1 cup of strawberries
1 cup of cut-up bananas
2 cups of whipped cream
1 cherry

Specifications for picture to be drawn:



Write a paragraph that explains to someone who has never fixed food how to put the Fruit and Whipped Cream Treat together.

- Start with the first thing you do.
- Tell about each part of the Treat that you put together.
- Tell about the parts in the order that you must put them together.
- Be sure to use exact words.
- Use words like next to join some of the sentences in your paragraph.

Figure 11. SWRL Writing Sample Stimulus.

Scoring Criteria	Good	Acceptable	Unacceptable
CONTENT:			
Includes all parts of the Treat.	___	___	___
Writes about the steps in the order in which they are performed.	___	___	___
Uses precise language.	___	___	___
Uses logical transitions.	___	___	___
Uses sentences that pertain to the main idea.	___	___	___
FORM:			
Uses correct grammar.	___	___	___
Uses good sentence structure.	___	___	___
Capitalizes correctly.	___	___	___
Punctuates correctly.	___	___	___
Spells correctly.	___	___	___
Writes legibly, with appropriate margins and indentation.	___	___	___

Figure 12. SWRL Scoring Key for Figure 11.

SWRL scoring key displayed in Figure 12 typically simplifies the evaluation of these characteristics by describing how they are specifically implemented for this individual task:

Trait	Implementation
Specific information:	Includes all parts of the treat.
	Uses sentences that pertain to the main idea (this criterion also assesses coherence).
Sequential presentation:	Writes about the steps in the order in which they are performed.
	Uses logical transitions (this criterion also assesses coherence).
Precise language:	Uses precise language.

This specificity contrasts sharply with the general criteria for evaluating an expository (explanatory) task in the analytic model displayed in Figure 2.

In contrast with the holistic system, SWRL scoring provides considerable diagnostic information. In fact, SWRL scoring provides more detailed diagnostic information than does PTS. For PTS, various aspects of content are collapsed into one category, the primary trait, and thus are given a single point score for the degree to which this overall trait is exhibited (see, for example, imaginative and evocative details, topic sentences, transitions, etc., displayed in the PTS guidelines in Figure 7). Thus PTS does not provide diagnostic information for individual components (e.g., evocative details, transitions, topic sentences) of the primary trait. When form skills such as grammar and mechanics are assessed as the secondary or tertiary trait

in PTS, these skills are also collapsed for a single score (e.g., the punctuation, capitalization, and other conventions as the tertiary trait in Figure 10), so that, for example, a score of two on the tertiary trait would not indicate whether errors reflected problems in punctuation, capitalization, or other conventions, or of a combination of these categories.

In contrast with PTS, the SWRL model lists content and form features for separate scoring, and performance ratings of good, acceptable, or unacceptable are given on each criterion. Thus the scoring elicits considerable diagnostic information about the individually listed component skills. Also in contrast with PTS (and consonant with analytic scales), the system provides a total score for the writing sample when numerical equivalents are assigned to the good/acceptable/unacceptable ratings of the content and form skills and these individual scores are tallied. Furthermore, the SWRL model does not prescribe the numerical values to be assigned to the good/acceptable/unacceptable categories as does the analytic scale; this "omission" allows users of the scoring system to assign weighted values that correspond to the importance of the skills to the population to be assessed.

As with the other systems, a scoring guide (see Figure 13) accompanies the stimulus; this guide describes the features that constitute a good, acceptable, or unacceptable score on each criterion in the scoring key. In contrast with PTS, these guidelines are simpler to interpret and thus to use because they correspond to precise individual features that implement the whole "primary trait." This simplicity is evident when the guidelines in Figures 13 and 7 are compared. In contrast

with analytic scales (and consonant with PTS), the guidelines are specific to the task described in the writing-sample stimulus rather than to any stimulus for a particular writing type. This specificity to the stimulus is evident when the guidelines in Figures 13 and 2 are compared.

Scoring is further simplified in the SWRL model by a range that spans three broad rating categories. Although authorities differ over the number of ratings a scoring range should span, they generally agree that the fewer the possible rankings, the simpler the scoring. Holistic scales may frequently entail as many as eight or nine points (A Common Ground, 1978) or as few as four (ETS, 1976), but these numbers are for an entire essay rather than for individual criteria.

At the other extreme from the holistic scoring range is the dichotomous scale. This scale consists of only two ratings for each feature/skill to be evaluated--acceptable/unacceptable or yes/no. As noted earlier, Cooper (1977) doubts that dichotomous scales yield reliable scores on individuals. Additionally, dichotomous scales tend to place a disproportionate number of students into the unacceptable range, and this phenomenon is inappropriate for competency testing. In an unpublished pilot study (Humes, 1979a), evaluators reported that decision making was difficult in dichotomous scoring since superior papers must be assigned the same acceptable/yes rating as mediocre papers. Because of the wide disparity in performance, scorers tended to rate borderline papers as unacceptable in comparison with superior papers.

A four-point scale is used by NAEP in Primary Trait Scoring: "The first and lowest score point is for respondents who show no evidence of the skill being measured. The second point is defined as some

Includes all parts of the Treat (i.e., plate, cake, strawberries, whipped cream, banana slices, whipped cream, cherry).

Good: All parts are included.

Acceptable: Most parts are included.

Unacceptable: Many parts are missing.

Writes about the steps in the order in which they are performed (i.e., cake on plate, strawberries on cake, whipped cream on strawberries, banana slices on whipped cream, whipped cream on banana slices, cherry on whipped cream/cherry on top).

Good: Most or all steps are in the right order.

Acceptable: Many steps are in the right order.

Unacceptable: Many steps are not in the right order.

Comment: If fewer than three steps are included, do not score this criterion.

Uses precise language.

Good: Precise words are generally used to explain the procedure; e.g., an exact word like "spread" is used to explain adding a food part to the top of a previous food part.

Acceptable: General words are used to explain the procedure; e.g., "put on" is used to explain adding a food part to the top of a previous food part.

Unacceptable: Language used does not convey a sense of the procedure.

Uses logical transitions.

Good: Several logical transitions for time (e.g., "then," "next"), space (e.g., "on top of"), and possibly reason (e.g., "because") are used.

Acceptable: Few logical transitions are used.

Unacceptable: No logical transitions are used.

Figure 13. Sample of SWRL Scoring Guidelines

evidence of the skill. Score point '3' is defined as demonstrating competence, and score point '4' is reserved for those papers which are excellent or outstanding" (Mullis, 1975, p. 10). Educational Testing Service (ETS) also uses a four-point scale because "In a scale which has only four points (4, 3, 2, 1), there is obviously no middle score" (1976, p. 3). ETS adds that "One of the major reasons why the scale works as well as it does is that readers are forced away from a middle, or uncommitted, score" (p. 3).

Forcing scores away from the middle range, as is done by both NAEP and ETS*, is inappropriate for minimal competency testing. A four-point scale implies that two is below average, and, inevitably, more writing samples will be perceived as below average because the scale has two categories below the midpoint of the range.

Five-point scales are sometimes used for analytic scoring in accordance with Diederich's model. Although the scales span five points, the corresponding guidelines describe only three ratings: low, middle, and high (Diederich, 1974). No guidelines are provided to aid scorers in discriminating between 1 and 2 or between 4 and 5. Further, Lloyd-Jones (1977) criticizes the five-point scale, asserting that "'3' is not the midpoint of a five-point scale; in this rating '5' is exceptionally good. '1' and '2' represent failures. . . '3' and '4' represent degrees of success. '5' is a rarity of success. . ." (p. 50).

For the reasons above, only a three-point scale was acceptable for the SWRL model. Because there is only one category below the midpoint, fewer papers are likely to be scored as below average with

*NAEP is not assessing for individual performance; much of ETS testing is for purposes other than minimal competency assessment.

this scale than with scales of two, four, or five. Additionally, the broad discriminations evaluators must make do not entail so much information processing as do systems with wider ranges: Wider ranges require evaluators to remember too many distinctions for the different ratings of all the features to be scored (e.g., in the SWRL model, a scoring key with ten content/skill categories and a span of three ratings requires 30 discriminations; in a five-point scale for the same number of content/skill categories, 50 discriminations are required). Furthermore, larger ranges may decrease both the reliability and validity of the competency scores.

A pilot study has provided evidence in support of the current design. Field testing of the SWRL model has produced data on the appropriateness of the design and the range of the scoring (Humes, 1979b). Scorers were given no prior training, yet results suggest that scoring of writing samples with this system is rapid and yields consistent rank-order scores. The mean time for reading and scoring papers at grade 2 was only 1.95 minutes, with a range of 1.17 to 2.67 across scorers; the mean time for the process with grade 6 papers was 2.8 minutes, with a range of 1.67 to 3.33. Although individual evaluators scored papers consistently higher or lower then did other scorers, the rank order of samples generally matched that of other evaluators

The careful structuring of the complete assessment package (i.e., stimulus, scoring key, and scoring guidelines) is an exacting and time-consuming task. However, great effort put into the development of the package results in minimal effort for users. Evaluating writing samples

is then a comparatively simple task that can be quickly expedited by the classroom teacher when the SWRL system is used. The model presented here is appropriate and efficient for eliciting and evaluating writing samples administered to determine students' writing competency. Furthermore, the diagnostic information elicited by this system can be used to ameliorate instructional inadequacies and assure student learning. Such results will go a long way toward allaying professional concern about competency testing.

REFERENCES

A common ground for assessing competence in written expression. Publication of the Office of the Los Angeles County Superintendent of Schools, Division of Curriculum and Instructional Services, 1978.

An assessment of the writing performance of California high school seniors. Sacramento: California State Department of Education, 1977.

Annich, S. V. The measurement of writing skills: A challenge for the test developer. Paper presented at AERA, San Francisco, April, 1979.

Brown, R. Insights from a decade of national assessment. Paper presented at AERA, San Francisco, April, 1979.

Cooper, C. R. Holistic evaluation of writing. In Cooper, C. R., and Odell, L., eds., Evaluating Writing: Describing, measuring, judging. Buffalo: State University of New York, 1977, pp. 3-31.

Diederich, P. B. How to measure growth in writing ability. English Journal, 1966, 55, pp. 435-449.

Diederich, P. B. Measuring growth in English. Urbana, Ill.: National Council of Teachers of English, 1974.

Educational Testing Service. How the essay in the CEEB English composition test is scored: An introduction to the reading for readers. Princeton: Educational Testing Service, 1976.

Friedman, M., & Fowler, E. Assessing elementary students' writing skills. Paper presented at AERA, San Francisco, April, 1979.

Hogan, T. P. & Mishler, C. J. Judging the quality of student's writing: When and how. The Elementary School Journal, 79:3, 1979, pp. 142-146.

Humes, A. A pilot study of dichotomous scoring of writing samples (1979a), unpublished.

Humes, A. Writing sample, scoring keys, and scoring guides: A pilot study (Technical Note No. 3-79-05). Los Alamitos, Calif.: SWRL Educational Research and Development, 1979.

Kadomoto, L. Hawaii writing assessment. Unpublished 1979 document received by personal communication with L. Kadomoto, Evaluation Specialist for the Department of Education for the State of Hawaii.

Klaus, C. H., Lloyd-Jones, R., Brown, R., Littlefair, W., Mullis, I., Miller, D., & Verity, D. Composing childhood experience: An approach to writing and learning in the elementary grades. St. Louis: CEMREL, Inc., 1979.

Lloyd-Jones, R. Primary trait scoring. In Cooper, C. R. & Odell, L., eds., Evaluating writing: Describing, measuring, judging. Buffalo: State University of New York, 1977, pp. 33-66.

Mullis, I. V. S. The primary trait system for scoring writing tasks. Denver: National Assessment of Educational Progress, 1975.

National Assessment of Educational Progress. Report 3: 1969-1970 writing: National results. Denver: Education Commission of the States, 1970.

Odell, L. & Cooper, C. Evaluating procedures for evaluating writing. Paper presented at AERA, Toronto, March, 1978.

Pitts, M. The measurement of students' writing performance in relation to instructional history. Paper presented at AERA, San Francisco, April, 1979.

Rentel, V. M., King, M., and Pappas, C. C. Toward a theory of early writing development. Paper presented at AERA, San Francisco, April, 1979.

Scherich, H. H., Liu, S., Kar, R., & Kettel, R. The development of writing production assessment items and scoring systems for the Louisiana State Pupil Assessment Program. Paper presented at AERA, San Francisco, April, 1979.

Smith, L. S. Measures of high school students' expository writing: Direct and indirect strategies. Paper presented at AERA, San Francisco, April, 1979.

Stahl, A. Structural analysis of children's compositions. Research in the Teaching of English, 1974, 8, pp. 184-205.

Winters, L. Alternative scoring systems for predicting criterion group membership. Paper presented at AERA, San Francisco, April, 1979.